



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Med Zipf mot framtiden En integrerad lexikonresurs för svensk språkteknologi

Lars Borin

Språkbanken
Inst. för svenska språket
Göteborgs universitet

Schæffergårdssymposiet 30/1 2010



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

bakgrund och förutsättningar

- ▶ lexikonresurser för språkteknologi kräver stora arbetsinsatser för sitt förverkligande
- ▶ resurserna blir användbarare om de är interoperabla än om de inte är det, även mellan språk
- ▶ en sorts mångsidigt användbar resurs är ett frasnät (framenet)
- ▶ frasnät finns för några få språk, men inte för svenska
- ▶ Språkbanken har startat ett svenskt frasnätsprojekt, SweFN++
- ▶ i det försöker vi återanvända befintliga fria resurser så mycket som möjligt



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

återanvändning av befintliga resurser

i Språkbanken

- ▶ SALDO (~73.000 betydelser) (navet i SweFN++)
- ▶ PAROLE (~29.000 lemgram) (syntaktisk valens)
- ▶ SIMPLE (~4000 betydelser) (semantisk valens)
- ▶ GLDB/SDB (~60.000 ingångar) (semantisk valens)
- ▶ Dalin (1800-t.; ~60,000 ingångar) (diakroni)
- ▶ (fornsvenska (~25.000 ingångar) (diakroni))
- ▶ (~200 miljoner ord korpusar)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SALDO

SALDO - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/sal/

Mest besökta clt Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by...

SALDO

SALDO

| lexem | mor | far | lemma |
|-------|---------|------|---------------------|
| snits | snitsig | PRIM | snits ⁿⁿ |

Ladda om sidan för nytt lexem.

[In English](#)

[Lars Borin, Markus Forsberg och Lennart Lönngrén](#)

Introduktion

SALDO (Svenskt associationslexikon 2) är en omfattande lexikonresurs för modernt svenskt skriftspråk som är avsedd för användning i språkteknologisk forskning och utveckling av språkteknologiska applikationer. Man kan betrakta SALDO som baslexikonresursen i en svensk BLARK. SALDO bygger på Svenskt associationslexikon, ett semantiskt lexikon för svenska.

SALDO är en elektronisk lexikonresurs avsedd för språkteknologiska tillämpningar. Det betyder att den är strukturerad i enlighet med detta och har ett innehåll som överensstämmer med detta mål, ett innehåll som av den anledningen på viktiga punkter avviker från vad man förväntar sig att finna i traditionella lexikon avsedda för mänskligt bruk (som ju också i allt större utsträckning görs tillgängliga i elektronisk form). Slutligen distribueras den i ett format som är avsett och lämpar sig för programmatisk åtkomst, alltså användning som en komponent i datorprogram.

Relaterade länkar

- [Sök i SALDO \(Dokumentation: Saldonstruktion.pdf\)](#)
- [Statistik, historik och felrapport](#)
- [SALDO:s nättjänster](#)

Klar

zotero



SALDO, 2

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

lexikon..1 - Mozilla Firefox

Arkiv Bedigera Visa Histgrik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/ws/saldo-ws/lid/html/lexikon..1

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen for ... Language Log The LINGUIST ... OREL - All links by...

lexikon..1

SALDO

skicka

| | |
|--------|--|
| lex: | lexikon (nn) |
| fm: | ordbok |
| fp: | PRIM |
| mf(5): | PRIM: lexikalisk ² association: associationslexikon ordkonstruktion: konstruktionslexikon rim: rimlexikon språkteknologi: SALDO |
| pf(0): | * |

Klar

zotero



SALDO, 3

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

PRIM..1 - Mozilla Firefox

Arkiv Bedigera Visa Histgrik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/ws/saldo-ws/lid/html/PRIM..1

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by...

PRIM..1

SALDO

skicka

| | |
|------------|---|
| lex: | PRIM (*) |
| fm: | * |
| fp: | * |
| mf(50): | PRIM: all annan använda att bara bra den fort framme färg för ² förbi före genom göra ha hur hända i ² ja just kunna ljud ljus med men mycken måste namn natur när och om om ² på rak röra säga tal till tänka vad var vara varm vem veta vid vilja öppen |
| pf(42334): | * |

Klar

zotero



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Dalins ordbok (1850–53)

DALINS ORDBOK (Språkbanken)> - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/dalin/ Dictionary.com

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by...

DALINS ORDBOK (Språkbanken)>

UPPSLAGSORD

gnagare

MAX ANTAL TRÄFFAR 10 SÖK

[Allt om Dalins ordbok](#)

GNAGARE , m. 5. En, som gnager. Brukas nästan endast i nat. hist. om djur, tillhörande 4:de ordningen af Däggdjuren eller dem, som sakna hörntänder, men deremot hafva två långa, hvassa framtänder, såsom t. ex. råttorna.

Hämtar data från sprakbanken.gu.se...



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

externa fria lexikonresurser

- ▶ SynLex – folkets synonymordbok (för 'variabelt' ordnät)
- ▶ svenska Wiktionary (för definitioner)
- ▶ Lund University Frame List (för frasnätsord)
- ▶ IDS/LWT
- ▶ (¿¿) svenskt ordnät (??)
- ▶ ... och kanske andra



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SynLex

Folkets synonymlexikon Synlex

Alla som sökte i svensk-engelska Lexin 24 mars-15 november 2005 fick möjlighet att bedöma ett slumpmässigt framtaget synonymparsförslag. Nu kan man slå i ett synonymlexikon med omkring 80000 synonympar som bedömts vara tillräckligt bra. Svara bara på [en synonymparsfråga](#) så får du chans att söka i Folkets synonymlexikon. Det går att [ladda hem synonymlexikonet](#) i XML-format. Eftersom Folkets synonymlexikon har byggts upp tillsammans av tiotusentals frivilliga Internetanvändare är lexikonet fritt.

Historik

Synonymlexikonet började byggas upp den 24 mars 2005. Då fanns 250000 synonymparsförslag. Den 7 april sållades 10% av synonymparsförslagen bort efter att många bedömt paren som synonymer av grad 0. Den 17 maj hade totalt 85000 synonymparsförslag sorterats bort, varefter 63% återstod. Kvaliteten på återstående synonymparsförslag blev då betydligt bättre. Från och med juni 2005 sållas även synonymparsförslag som ansetts bra bort från de synonympar som presenteras för bedömning, men de är förstas kvar i Folkets synonymlexikon. Ungefär 25000 användarsynonymförslag har lagts till. Även om vi försöker ta bort felstavade och tokiga förslag så kan en del förslag som presenteras vara oseriösa. Vi ber användarna att bortse från dessa förslag och helt enkelt svara 0 på synonymitetsfrågan.

Vad menar vi med att två ord är synonymer?

Synonymer är ord som betyder samma sak. Men det är mycket ovanligt med ord som är helt utbytbara. Vissa ord har till exempel olika stilvärde (flicka och tös). Andra har överlappande men inte identisk betydelse (god och smaklig). Detta gör att språkvetare är försiktiga med att definiera exakt vilka ord som är synonymer.

I Synlex är det folkets definition av synonym som används. Det betyder att om tillräckligt många personer anser att två ord är synonymer (till en viss grad) så kommer orden att vara det i lexikonet.

Om synonymparsförslagen

Vissa förslag kan tyckas vara helt gripna ur luften. Ibland föreslås ord ur olika ordklasser som synonymförslag. Hur är det möjligt?

Klar



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

IDS (Intercontinental Dictionary Series)


IDS Project - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://lingweb.eva.mpg.de/ids/ Dictionary.com

Mest besökta Welcome to CU ... Google Lingvistbloggen Institutionen for ... Language Log The LINGUIST ... OREL - All links by...

IDS Project



[IDS Main Page](#)
[Simple Browsing](#)
[Advanced Browsing](#)
[Download Data](#)
[Technical](#)
[Background](#)

The Intercontinental Dictionary Series

Founding Editor:
Mary Ritchie Key (University of California, Irvine)

General Editor:
Bernard Comrie (Max Planck Institute for Evolutionary Anthropology, Leipzig)

Purpose: The purpose of the IDS is to establish a database where lexical material across the continents is organized in such a way that comparisons can be made. Historical studies, comparative, and theoretical linguistic research can be based on this documentation. This is a long term cooperative project that will go on for the next generation or so and will involve linguists all over the world. It is aimed towards international understanding and cooperation. This is a pioneering effort that will have global impact. The purpose also contributes to preserving information on the little-known and "non-prestigious" languages of the world, many of which are becoming extinct.

Rationale: Information on languages of the world is scattered over all the continents and islands and published in dozens of languages and scripts. There is need of a database where one can find comparable material to formulate hypotheses and test and validate those theories. For example, theories on intercontinental connections have been proposed on the basis of the distribution of 'sweet potato' and yet there is no single source, where words with this meaning can be found in many languages. Good quantitative and statistical studies are almost impossible to do now in non-Western languages. The IDS will provide a quantitative base for a scientific approach to language analysis and comparisons. The IDS will provide the research tools necessary for expanding studies such as phonological theory, word formation, language change, lexical distribution,

Klar

zotero



IDS, 2

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT


IDS Project - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://lingweb.eva.mpg.de/ids/

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by...

IDS Project



IDS Main Page
Simple Browsing
Advanced Browsing
Download Data
Technical
Background

klär

- ☐ Casino [Phonemic]
- ☐ Catalan [StandardOrth]
- ☐ Catuquina [Phonemic]
- ☐ Cavineña [Phonemic]
- ☐ Cayapa [Phonemic]
- ☐ Cayuvava [Phonemic]
- ☐ Chamalal [CyrillTrans]
- ☐ Chatino, Zacatepec [Phonemic]
- ☐ Chácobo [Phonemic]
- ☐ Chechen [CyrillTrans]
- ☐ Chechen Dialect Akkin [CyrillTrans]
- ☐ Chehalis (Upper) [Phonemic]
- ☐ Chipaya [Phonemic]
- ☐ Chiriguano [Phonemic]
- ☐ Chorote [Phonemic]
- ☐ Cofán [Phonemic]
- ☐ Colorado [Phonemic]
- ☐ Czech [StandardOrth]
- ☐ Danish [StandardOrth]
- ☐ Dargwa [CyrillTrans]
- ☐ Dargwa Dialect Kajtak [CyrillTrans]
- ☐ Dargwa Dialect Kubachi [CyrillTrans]
- ☐ Dargwa Dialect Muiri [CyrillTrans]
- ☐ Dargwa Itsari [CyrillTrans]
- ☐ Dutch [StandardOrth]
- ☐ Elamite [Phonemic]
- ☐ Embera [Phonemic]
- ☐ English (StandardOrth)
- ☐ English (Middle) [StandardOrth]
- ☐ English (Old) [StandardOrth]
- ☐ Eöena [Phonemic]
- ☐ Fula [Phonemic]
- ☐ Polish [StandardOrth]
- ☐ Portuguese [StandardOrth]
- ☐ Proto Austronesian [Phonemic]
- ☐ Proto Polynesian [Phonemic]
- ☐ Prussian (Old) [StandardOrth]
- ☐ Puinave [Phonemic]
- ☐ Qawasqar [Phonemic]
- ☐ Rapa Nui [Phonemic]
- ☐ Rapa Nui [Phonemic (vars)]
- ☐ Romani [Phonemic]
- ☐ Romanian [StandardOrth]
- ☐ Rotuman [Phonemic]
- ☐ Russian [Phonemic]
- ☐ Rutul [CyrillTrans]
- ☐ Sanapaná Dialect Angaité [Phonemic]
- ☐ Sanapaná Dialect Enlhet [Phonemic]
- ☐ Sanskrit [LatinTrans]
- ☐ Selknam [Phonemic]
- ☐ Selkup [Phonemic]
- ☐ Serbo-Croatian [Phonemic]
- ☐ Seri [Phonemic]
- ☐ Shipibo-Conibo [Phonemic]
- ☐ Siona [Phonemic]
- ☐ Siriono [Phonemic]
- ☐ **Other Names**
- ☐ SIL: Swedish [Phonemic]
- ☐ SIL: Swedish [Phonemic]
- ☐ Swensen [StandardOrth]
- ☐ Tabassaran Dialect North Tabasaran (Khanag) [CyrillTrans]
- ☐ Tabassaran Dialect South Tabasaran [CyrillTrans]

zotero



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

LWT (Loanword Typology Project)

Mozilla Firefox

Arkiv Bedigera Visa Histgrik Bokmärken Verktyg Hjälp

http://email.eva.mpg.de/~haspelmt/lwt-meanings.htm Dictionary.com

Mest besökta Welcome to CU ... Google Lingvistbloggen Institutionen for ... Language Log The LINGUIST ... OREL - All links by...

Languages - Loanword Typol... http://email.e...-meanings.htm

The Loanword Typology Meaning List

This list is used by the [Loanword Typology project](#) and is based on the list of the [Intercontinental Dictionary Series](#). The meaning descriptions and the typical contexts were added by Martin Haspelmath and Uri Tadmor.

- 1 [The physical world](#) |
- 2 [Kinship](#) |
- 3 [Animals](#) |
- 4 [The body](#) |
- 5 [Food and drink](#) |
- 6 [Clothing and grooming](#) |
- 7 [The house](#) |
- 8 [Agriculture and vegetation](#) |
- 9 [Basic actions and technology](#) |
- 10 [Motion](#) |
- 11 [Possession](#) |
- 12 [Spatial relations](#) |
- 13 [Quantity](#) |
- 14 [Time](#) |
- 15 [Sense perception](#) |
- 16 [Emotions and values](#) |
- 17 [Cognition](#) |
- 18 [Speech and language](#) |
- 19 [Social and political relations](#) |
- 20 [Warfare and hunting](#) |
- 21 [Law](#) |
- 22 [Religion and belief](#) |
- 23 [Modern world](#) |
- 24 [Miscellaneous function words](#)

Semantic field 1: The physical world

Klar zotero



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

hopkoppling och harmonisering av lexikonresurser

- ▶ befintliga resurser i flera olika format, med olika sorters innehållskategorier
- ▶ minimimålet är två sorters gemensamma enheter:
 1. betydelser
 2. lemgram (och tillhörande böjningsmönster)
- ▶ vi vill kunna koppla ihop båda sorterna över alla resurser
- ▶ all information måste vara **explicit** och **entydig**
- ▶ vi vill använda SALDO:s identifierare



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

stabla identifierare ("PID")

- ▶ i SALDO finns id för:
 - ▶ betydelser (*grad..1*)
 - ▶ lemgram (*grad..nn.1*)
 - ▶ ordklasser (*nn*)
 - ▶ paradigm/böjningstabeller (*nn_3u_film*)
- ▶ SALDO-id:na är utformade för att vara:
 - ▶ **unika** (inga andra id behövs, t.ex. som databasnycklar)
 - ▶ **atomära** (inga inbyggda antaganden om betydelser/underbetydelser, etc.) (fast lemgram-id innehåller ordklassbeteckningen)
 - ▶ **användbara i Semantic Web-formalismer** (RDF, OWL): de är välformade xml-namn
 - ▶ **läsbara av människor** (underlättar arbetet med resurserna)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

hopkoppling av lexikonresurser

- ▶ harmoniseringen av format kan i stor utsträckning göras med datorprogram
- ▶ hopkopplingen av resursernas innehåll vill vi också automatisera så mycket som möjligt
- ▶ men hur mycket är möjligt? hur mycket manuellt arbete kommer det att innebära?
- ▶ det är här **Zipf** kommer in i bilden



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

George Kingsley Zipf (1902–1950)



(från Wikipedia)

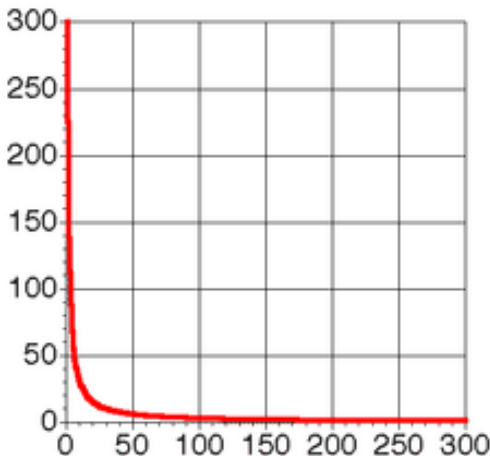


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Zipfs lag: rangordning \Rightarrow frekvens



(från <http://www.useit.com>)

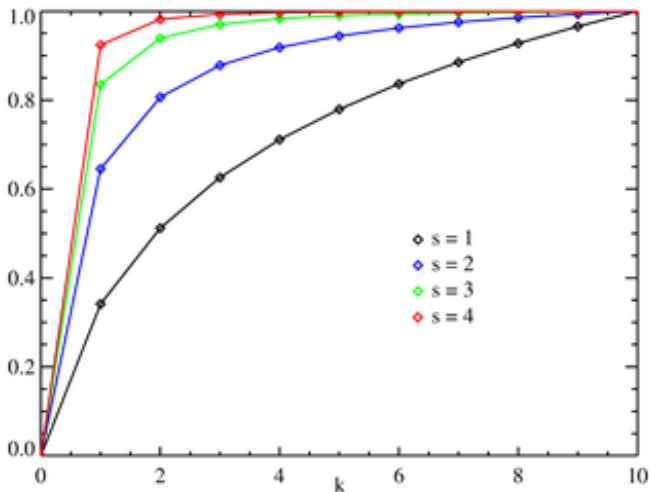


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Zipfs lag: rang \Rightarrow kumulativ mängd



(från Wikipedia)

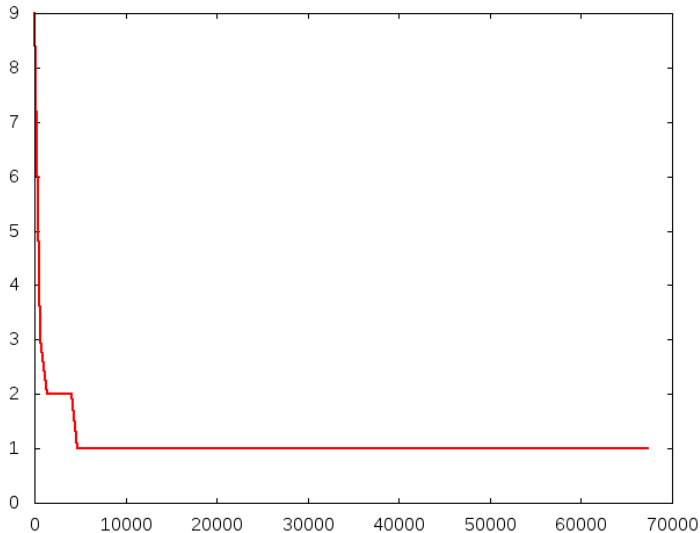


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

betydelser per grundform i SALDO



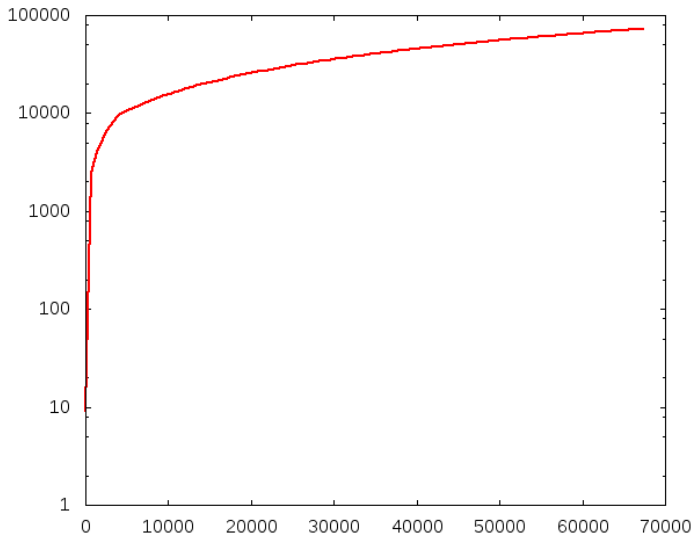


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

grundformer \Rightarrow betydelser i SALDO





GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

med Zipf . . .

- ▶ mot framtiden
- ▶ och mot det förflutna
- ▶ hypotes: eftersom de flesta grundformerna bara bär en betydelse i våra lexikonresurser, kan vi mekaniskt koppla ihop dem med acceptabel precision för praktiska tillämpningar

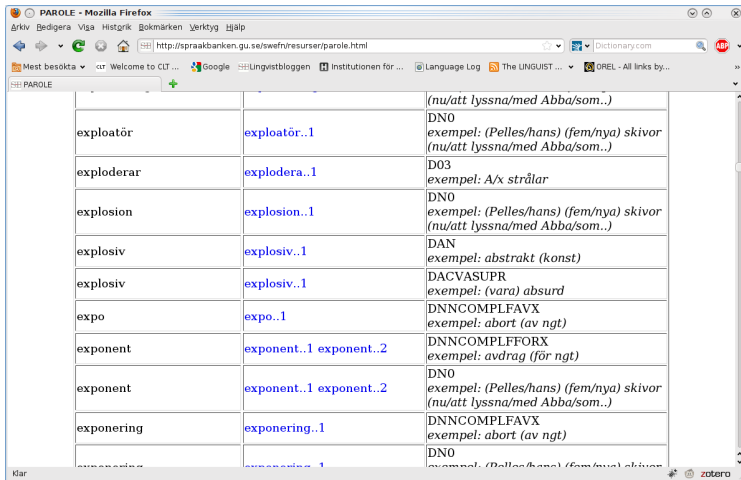


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

PAROLE – SALDO



| | | | |
|------------|-------------------------|---------------|--|
| | | | (nu/att lyssna/med Abba/som..) |
| exploatör | exploatör..1 | DN0 | exempel: (Pelles/hans) (fem/nya) skivor (nu/att lyssna/med Abba/som..) |
| exploderar | explodera..1 | D03 | exempel: A/x strålar |
| explosion | explosion..1 | DN0 | exempel: (Pelles/hans) (fem/nya) skivor (nu/att lyssna/med Abba/som..) |
| explosiv | explosiv..1 | DAN | exempel: abstrakt (konst) |
| explosiv | explosiv..1 | DACVASUPR | exempel: (vara) absurd |
| expo | expo..1 | DNNCOMPLFAVX | exempel: abort (av ngt) |
| exponent | exponent..1 exponent..2 | DNNCOMPLFFORX | exempel: avdrag (för ngt) |
| exponent | exponent..1 exponent..2 | DN0 | exempel: (Pelles/hans) (fem/nya) skivor (nu/att lyssna/med Abba/som..) |
| exponering | exponering..1 | DNNCOMPLFAVX | exempel: abort (av ngt) |
| exponering | exponering..1 | DN0 | exempel: (Pelles/hans) (fem/nya) skivor |



SIMPLE – SALDO

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SIMPLE - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/swefw/resurser/simple_v.html

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by...

SIMPLE

| | | | | | |
|-------------|---------------|--------|----|----------------------------|--|
| | Communication | x1/1/0 | p1 | a0 | a0-t-Human-tra |
| löddrar | <<2/1 | =2/1/0 | ZZ | +State | Gen |
| | Stative | x1/1 | p1 | a0 | a0-t-Substance-tra |
| löder | <<1 | =1/1/0 | ZZ | +Cause_constitutive_change | Gen |
| | Change | x1/1 | p1 | a0 a1 | a0-t-Human-tra a1-t-Concrete_entity-tra |
| lokaliserar | <<1 | =1/1/0 | ZZ | +Acquire_knowledge | Gen |
| | Cognition | x1/2/0 | p1 | a0 a1 | a0-t-Human-tra a1-t-Entity-tra |
| löper | <<1 | =1/2/0 | ZZ | +Stative_location | Gen |
| | Stative | x1/1 | p1 | a0 a1 | a0-t-Entity-tra a1-t-Location-tra |
| löper | <<2 | =1/4/0 | ZZ | +State | Gen |
| | Stative | x1/1 | p1 | a0 a1 | a0-t-Human-tra a1-t-Time-tra |
| lossnar | <<1 | =1/1/0 | ZZ | +Change_of_state | Gen |
| | Change | x1/1 | p1 | a0 | a0-t-Concrete_entity-tra |
| luckrar | <<1 | =1/1/0 | ZZ | +Cause_change_of_state | Gen |
| | Change | x1/1 | p1 | a0 a1 | a0-t-Human-tra a1-t-Area-tra |
| lufsar | <<1 | =1/1/0 | ZZ | +Move | Gen |
| | Motion | x1/1 | p1 | a0 a1 | a0-t-Human-tra a1-t-Location-dfa |
| luftar | <<1 | =1/1/0 | ZZ | +Purpose_act | Gen |

Klar

zotero

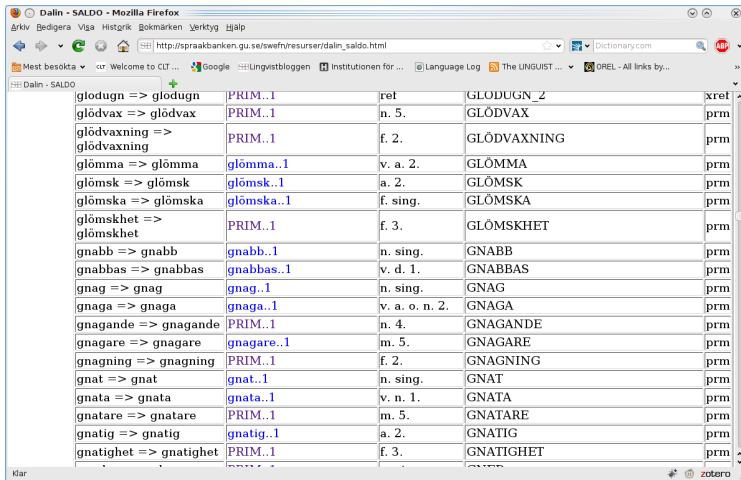


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Dalin – SALDO



| | | | | |
|-------------------------------|------------|----------------|-------------|------|
| glodugn => glödugn | PRIM..1 | ret | GLÖDUGN_Z | xret |
| glödvax => glödvax | PRIM..1 | n. 5. | GLÖDVAX | prm |
| glödvaxning => glödvaxning | PRIM..1 | f. 2. | GLÖDVAXNING | prm |
| glömma => glömma | glömma..1 | v. a. 2. | GLÖMMA | prm |
| glömsk => glömsk | glömsk..1 | a. 2. | GLÖMSK | prm |
| glömska => glömska | glömska..1 | f. sing. | GLÖMSKA | prm |
| glömskhet => glömskhet | PRIM..1 | f. 3. | GLÖMSKHET | prm |
| gnabb => gnabb | gnabb..1 | n. sing. | GNABB | prm |
| gnabbas => gnabbas | gnabbas..1 | v. d. 1. | GNABBAS | prm |
| gnag => gnag | gnag..1 | n. sing. | GNAG | prm |
| gnaga => gnaga | gnaga..1 | v. a. o. n. 2. | GNAGA | prm |
| gnagande => gnagande | PRIM..1 | n. 4. | GNAGANDE | prm |
| gnagare => gnagare | gnagare..1 | m. 5. | GNAGARE | prm |
| gnagning => gnagning | PRIM..1 | f. 2. | GNAGNING | prm |
| gnat => gnat | gnat..1 | n. sing. | GNAT | prm |
| gnata => gnata | gnata..1 | v. n. 1. | GNATA | prm |
| gnatare => gnatare | PRIM..1 | m. 5. | GNATARE | prm |
| gnatig => gnatig | gnatig..1 | a. 2. | GNATIG | prm |
| gnatighet => gnatighet | PRIM..1 | f. 3. | GNATIGHET | prm |



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SynLex – SALDO

| | |
|-----------|---|
| max: 92 % | |
| avg: 80 % | |
| dev: 0 % | fräsch..1 färsk..1 |
| min: 80 % | |
| max: 80 % | |
| avg: 80 % | |
| dev: 0 % | fröjda_sig..1 förlusta_sig..1 |
| min: 80 % | |
| max: 80 % | |
| avg: 80 % | |
| dev: 0 % | fukta..1 vattna..1 |
| min: 80 % | |
| max: 80 % | |
| avg: 85 % | |
| dev: 6 % | stinn..1 fylld..1 fullmatad..1 fullproppad..1 |
| min: 80 % | |
| max: 94 % | |
| avg: 86 % | |
| dev: 6 % | tänka_efter..1 grunna..1 grubbla..1 betänka..1 fundera..1 |
| min: 80 % | filosofera..1 tänka..1 |
| max: 96 % | |
| avg: 82 % | |
| dev: 0 % | fundering..1 tanke..1 |
| min: 82 % | |
| max: 82 % | |



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SynLex – SALDO, 2

| | |
|--|--|
| min: 90 % max: 90 % | tröjd..1 trivsel..1 |
| avg: 94 % dev: 0 % min: 94 % max: 94 % | fullproppad..1 stinn..1 |
| avg: 96 % dev: 4 % min: 92 % max: 100 % | fundamental..1 elementär..1 grundläggande..1 |
| avg: 94 % dev: 2 % min: 92 % max: 96 % | tänka_efter..1 fundera..1 tänka..1 |
| avg: 100 % dev: 0 % min: 100 % max: 100 % | klaffa..1 fungera..1 funka..1 |
| avg: 100 % dev: 0 % min: 100 % max: 100 % | furste..1 hövding..1 |
| avg: 90 % dev: 0 % min: 90 % max: 90 % | fyrhörning..1 romb..1 |



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SynLex – SALDO, 3

| | |
|--|--|
| min: 80 % max: 94 % | stinn..1 tynd..1 tuimatatad..1 tuuproppad..1 |
| avg: 85 % dev: 9 % min: 70 % max: 100 % | tänka_after..1 grunna..1 grubbla..1 begrundan..1 eftertanke..1 betänka..1 reflexion..1 begrunda..1 fundera..1 filosofera..1 tänka..1 |
| avg: 82 % dev: 0 % min: 82 % max: 82 % | fundering..1 tanke..1 |
| avg: 75 % dev: 5 % min: 70 % max: 80 % | tankfull..1 betänksam..1 fundersam..1 |
| avg: 76 % dev: 0 % min: 76 % max: 76 % | fura..1 tall..1 |
| avg: 80 % dev: 0 % min: 80 % max: 80 % | fy..1 usch..1 |
| avg: 82 % dev: 4 % min: 80 % max: 80 % | fyllo..1 suput..1 fyllerist..1 alkoholist..1 alkoholmissbrukare..1 |



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SynLex – SALDO, 4

| | |
|--|-------------------------------------|
| avg: 50 % dev: 0 % min: 90 % max: 90 % | friherre..1 ädling..1 |
| avg: 78 % dev: 8 % min: 70 % max: 86 % | frimärke..1 brevporto..1 porto..1 |
| avg: 90 % dev: 0 % min: 90 % max: 90 % | friskus..1 hurtbulle..1 |
| avg: 98 % dev: 2 % min: 96 % max: 100 % | hårfrisör..1 barberare..1 frisör..1 |
| avg: 86 % dev: 0 % min: 86 % max: 86 % | fru..1 hustru..1 |
| avg: 80 % dev: 0 % min: 80 % max: 80 % | frugal..1 sparsam..1 |
| | |



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SweFN-pilotprojekt

- ▶ personer: Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Annika Kjellandsson, Dimitrios Kokkinakis
- ▶ vi har använt existerande standardprogram för snabb start och utprövning av metodologi: Subversion, OpenOffice DB, emacs
- ▶ små specialprogram och -skript samt SALDO-webbtjänster (baserade på SALDO/FM-maskineriet) som klister
- ▶ återkoppling via automatgenererade webbsidor



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SALDO-återkoppling

SALDOs historik - Mozilla Firefox

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://demo.spraakdata.gu.se/markus/historik.html

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen for ... Language Log The LINGUIST ... OREL - All links by...

SALDO SALDOs historik SALDOs felrapport

SALDO

Nedan listas de senaste ändringarna i SALDOs grundmaterial.

Bla farg motsvarar tillägg och rød farg borttagning.

| | | | | | | |
|---------------|--------------|-------------|-------------------|------------|-----|----------|
| rep..1 | binda..1 | PRIM..1 | rep..nn.1 | rep | nn | nn_6n_b |
| rep..1 | binda..1 | PRIM..1 | rep..nn.1 | rep | nn | nn_6n_b |
| rep..2 | repa..4 | PRIM..1 | rep..nn.1 | rep | nn | nn_6n_b |
| repa..5 | reparera..1 | PRIM..1 | repa..vb.1 | repa | vb | vb_1a_l |
| analog..2 | digital..1 | motsats..1 | analog..av.1 | analog | av | av_1_g |
| boa_in_sig..1 | hem..1 | PRIM..1 | boa_in_sig..vbm.1 | boa in sig | vbm | vbm_1mps |
| digital..1 | sifra..1 | PRIM..1 | digital..av.1 | digital | av | av_1_g |
| digital..1 | dator..1 | PRIM..1 | digital..av.1 | digital | av | av_1_g |
| flyg..1 | flygplan..1 | PRIM..1 | flyg..nn.1 | flyg | nn | nn_6n_b |
| flyg..1 | flygplan..1 | PRIM..1 | flyg..nn.1 | flyg | nn | nn_6n_b |
| flyga..2 | resa..1 | flygplan..1 | flyga..vb.1 | flyga | vb | vb_4a_fl |
| flygförare..1 | flygplan..1 | PRIM..1 | flygförare..nn.1 | flygförare | nn | nn_6u_ki |
| pimpla..2 | dricka..1 | PRIM..1 | pimpla..vb.1 | pimpla | vb | vb_1a_vi |
| hotta..1 | hotta_opp..1 | PRIM..1 | hotta..vb.1 | hotta | vb | vb_1a_l |
| hotta..2 | hot..2 | PRIM..1 | hotta..vb.1 | hotta | vb | vb_1a_l |
| hottande..1 | hotta..1 | PRIM..1 | hottande..nn.1 | hottande | nn | nn_5n_s |
| hottande..2 | hotta..2 | PRIM..1 | hottande..nn.1 | hottande | nn | nn_5n_s |

Klar

zotero

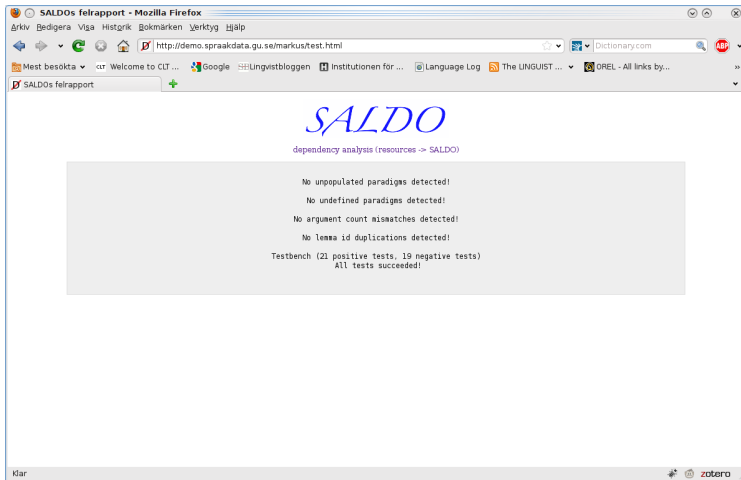


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SALDO-återkoppling, 2





GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

beroendekontroll gentemot SALDO

Beroendekontroll - Mozilla Firefox

Arkiv Bedigera Visa Histgrik Bokmärken Verktyg Hjälp

http://spraakbanken.gu.se/swefn/resurser/ref_report.html

Dictionary.com

Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen for ... Language Log The LINGUIST ... OREL - All links by...

Beroendekontroll

SweFN++

| | |
|------------------|-----|
| SweFN++ | OK! |
| Dalin | OK! |
| Swesaurus | OK! |
| LWT | OK! |
| Parole | OK! |

Klar

zotero



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SweFN++

The screenshot shows a web browser window titled "Svenskt frasnät (SweFN++) - Mozilla Firefox". The address bar shows the URL "http://spraakbanken.gu.se/swefn/". The browser's search bar contains "Dictionary.com". The page content features the "SweFN++" logo in a stylized green font, with "in English" written below it. The main heading is "Svenskt frasnät (SweFN++)". The text explains that the resource provides access to lexical, grammatical, and semantic information for Swedish, useful for automatic understanding and generation of natural language. It mentions that while similar resources exist for other languages, this one is specifically for Swedish. A project at Göteborgs universitet is noted for creating this resource. The "Utvecklingsversionen" (Development version) section lists links to HTML, XML, XSD, CSV files, a work report, a felrapport, a Berondeanalysis, and a history. The "Licens" (License) section states the resource is under the LGPL 3.0 or Creative Commons Attribution-Share Alike 2.5 Generic license. The footer shows the word "Lärare" and a Zotero icon.

Svenskt frasnät (SweFN++)

Tillgången till flera nivåer av lexikal, grammatisk och semantisk information som representerar textinnehåll är ett förkrav till automatisk förståelse och generering av naturligt språk. Ett frasnät anses vara en värdefull resurs för både lingvister och språkteknologer som kan bidra till att nå dessa mål.

För närvarande existerar frasnätliknande resurser endast för ett fåtal språk, se [Berkeleys FrameNet](#). Svenska är inte en av dem, även om vissa pilotundersökningar har gjorts för att halvautomatiskt skapa svenska frasnätsramar.

På Göteborgs universitet har vi nu påbörjat ett projekt att skapa ett svenskt frasnät. Något som utmärker detta projekt är att det svenska frasnätet kommer att bli en integrerad del av en större, mångfacetterad lexikal resurs. Därav namnet SweFN++.

Utvecklingsversionen

- Utvecklingsversionen av det svenska frasnätet: [HTML](#) [XML](#) [\(XSD\)](#) [CSV](#)
- Arbetsrapport för svenskt frasnät: [swefn-dbdok.pdf](#)
- [Felrapport](#)
- [Berondeanalysis](#) (resurser -> SALDO)
- Historik: [dd mtg](#)

Licens

Resursen sprids under friprogramvarulicensen [LGPL 3.0](#) eller [Creative Commons Attribution-Share Alike 2.5 Generic](#).

Lärare



SweFN++, 2

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

| | |
|---|--|
| Svenska Framenet++ - Mozilla Firefox | |
| Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp | |
| http://spraakbanken.gu.se/swefn/resurser/swefn-db.html | |
| Mest besökta Welcome to CLT ... Google Lingvistbloggen Institutionen för ... Language Log The LINGUIST ... OREL - All links by... | |
| Svenska Framenet++ | |
| ram | Commerce_buy |
| exempel | Jag köpte örhängen från HM för 50 spann. |
| semantisk typ | 0//Transaction |
| domän | Gen |
| kärnelement | Buyer, Goods |
| exempel | När man köper leksaker till barn under tre år behöver man vara extra uppmärksam. Jag har köpt en resa med mitt kreditkort. Byxorna fyndade jag på Btk. |
| periferielement | Duration, Manner, Means, Money, Place, Purpose, Purpose_of_goods, Rate, Reason, Recipient, Seller, Time, Unit |
| sms | Buyer+LU, Goods+LU, Manner+LU, Means+LU, Purpose+LU, Time+LU |
| sms-exempel | Buyer+LU_EX konsumentköp Goods+LU_EX bilköp, fastighetsköp, lösoreköp Manner+LU_EX impulsköp Means+LU_EX kassaköp Purpose+LU_EX tröstköp, täckningsköp Purpose_of_goods+LU_EX sexköp Time+LU_EX förköp |
| saldo | vb: fynda..1 friköpa..1 inköpa..1 köpa..1 tillhandla_sig..1 nn: bilköp..1 brudköp..1 fastighetsköp..1 friköp..1 förköp..1 inköp..1 kassaköp..1 lösoreköp..1 köp..1 samköp..1 täckningsköp..1 uppköp..1 återköp..1 |
| saldo (nya) | stödköp..1 konsumentköp..1 motköp..1 tröstköp..1 sexköp..1 |

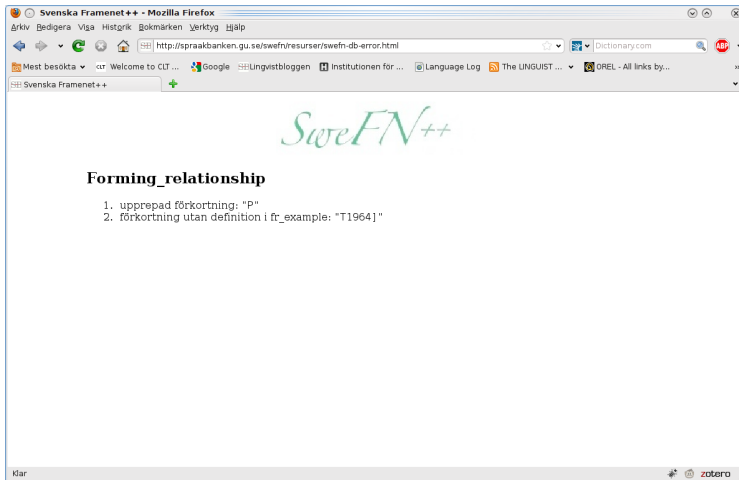


GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SweFN++, 3





GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

pågående och planerat arbete

- ▶ SALDO-städning (betydelser): nästan klart
- ▶ hopkoppling av SALDO och SynLex: påbörjad
- ▶ detaljerad inventering av fria lexikonresurser: gjord, men inte rapporterad
- ▶ pilotuppsättning frasnätsingångar: klar
- ▶ formatstandardisering (LMF/OWL) och innehållsharmonisering: våren 2010
- ▶ länkning SALDO – Dalin: våren 2010
- ▶ Dalinmorfologi: våren 2010
- ▶ medelsansökan: 10/2 (RJ), 30/3, 20/4 (VR)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

metodologi

- ▶ utgå från engelska ramar, modifiera vid behov (samma som SFN)
- ▶ använd de befintliga resurserna så mycket som det går (beroende av god länkning)
- ▶ hitta nya enheter i svenska korpusar



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

metodologi, 2

- ▶ använd befintliga verktyg för att annotera korpusar så att bra exempelfraser blir lätta att hitta (jfr Deepdict):
 - ▶ MALTParser kan ge (kandidater till) valensramar
 - ▶ SALDO (och annan lexikalisk-semantisk information) kan ge (kandidater till) semantiska grupper av valensargument
- ▶ något som vi är särskilt intresserade av är att utöka mängden flerordningar i våra lexikonresurser



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

metodologi, 3

- ▶ om arbetet läggs upp på rätt sätt, med rätt datorstöd, så hoppas vi att ny information i lexikonet direkt kan komma korpussökningen till nytta
- ▶ metodologiska och andra forskningsfrågor:
 - ▶ arbetsflöde och verktyg: hur bäst kombinera automatiska processer och manuellt arbete?
 - ▶ semantiska roller och ramelement
 - ▶ ramar och konstruktioner



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

sammanfattning

- ▶ planerna för SweFN++ omfattar:
 - ▶ återanvändning av fria lexikonresurser för svenska
 - ▶ byggande av ett svenskt frasnät ovanpå resurserna
 - ▶ skapandet av en diakronisk lexikonresurs
 - ▶ utforskande av metodologi för att göra allt detta så arbetsbesparande som möjligt
 - ▶ tillhandahållande av resultatet som en öppen resurs (open content/open source)